Introdução à Ciência de Dados

Cayan Portela

UniCEUB

March 3, 2023

O que é Ciência de Dados?

- Definição

Amazon Web Services (AWS)

"Uma abordagem multidisciplinar que combina princípios e práticas das áreas de matemática, estatística, inteligência artificial e engenharia da computação para analisar grandes quantidades de dados."

- Analisar dados
- Detecção de padrões
- Extrair informação

Onde estão os dados?

- Diversas fontes de dados
  - → Operações financeiras.
  - → e-commerce, web data.
  - → Repositórios de documentos.
  - → Redes sociais.

Onde estão os dados?

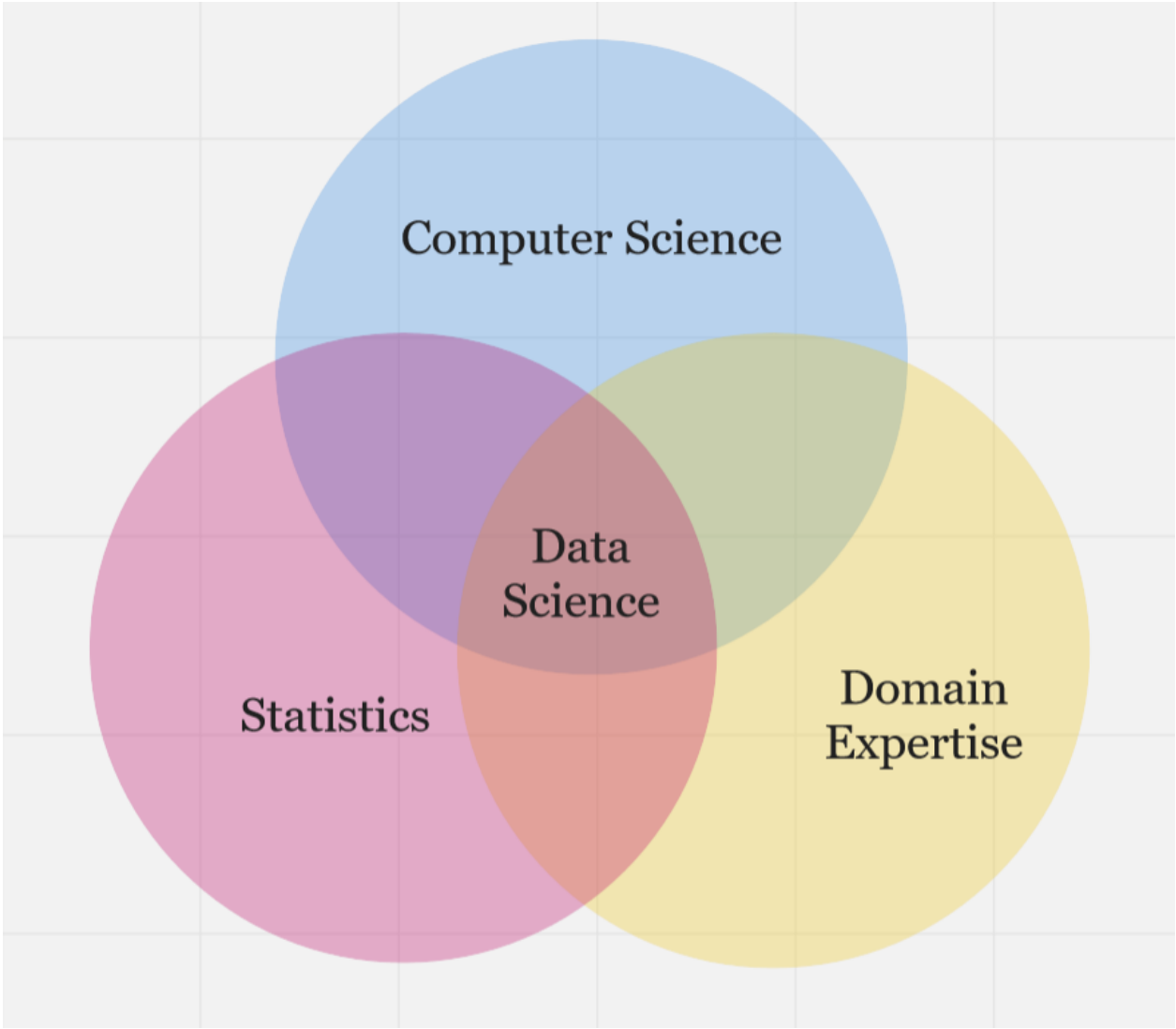
- Diversas fontes de dados
  - → Operações financeiras.
  - → e-commerce, web data.
  - → Repositórios de documentos.
  - → Redes sociais.

- Como estão sendo utilizados?
  - → Otimização de recursos.
  - → Recomendação.
  - → Tomadas de decisão.
  - → Personalização.

Aplicações

- Exemplos de casos de uso
  - → Aprovar ou não um empréstimo.
  - → Ofertas de produtos.
  - → Engajamento de usuários.
  - → Teste AB.
  - → Jurimetria.

Ciência de Dados



Ciência de Dados

- Matemática e Estatística
  - → Cálculo e Álgebra Linear (como otimizar uma função?).
  - → Probabilidade e Estatística.
  - → Inferência e Modelagem preditiva.

Ciência de Dados

- Matemática e Estatística
  - → Cálculo e Álgebra Linear (como otimizar uma função?).
  - → Probabilidade e Estatística.
  - → Inferência e Modelagem preditiva.

- Computação
  - → Cloud (GCP, AWS, Azure, etc)
  - → Linguagens de programação: Python, R, SQL.
  - → Ferramentas de versionamento (trabalho em grupo)
  - → Pipelines e steps de produtização.

Ciência de Dados

- Matemática e Estatística
  - → Cálculo e Álgebra Linear (como otimizar uma função?).
  - → Probabilidade e Estatística.
  - → Inferência e Modelagem preditiva.

- Computação
  - → Cloud (GCP, AWS, Azure, etc)
  - → Linguagens de programação: Python, R, SQL.
  - → Ferramentas de versionamento (trabalho em grupo)
  - → Pipelines e steps de produtização.

- Expertise na Área
  - → Qual métrica avaliar?
  - → Qual alteração realizar?
  - → Trade-off entre complexidade e resultados.
  - → Entendimento de todo o ecossistema do produto/serviço.

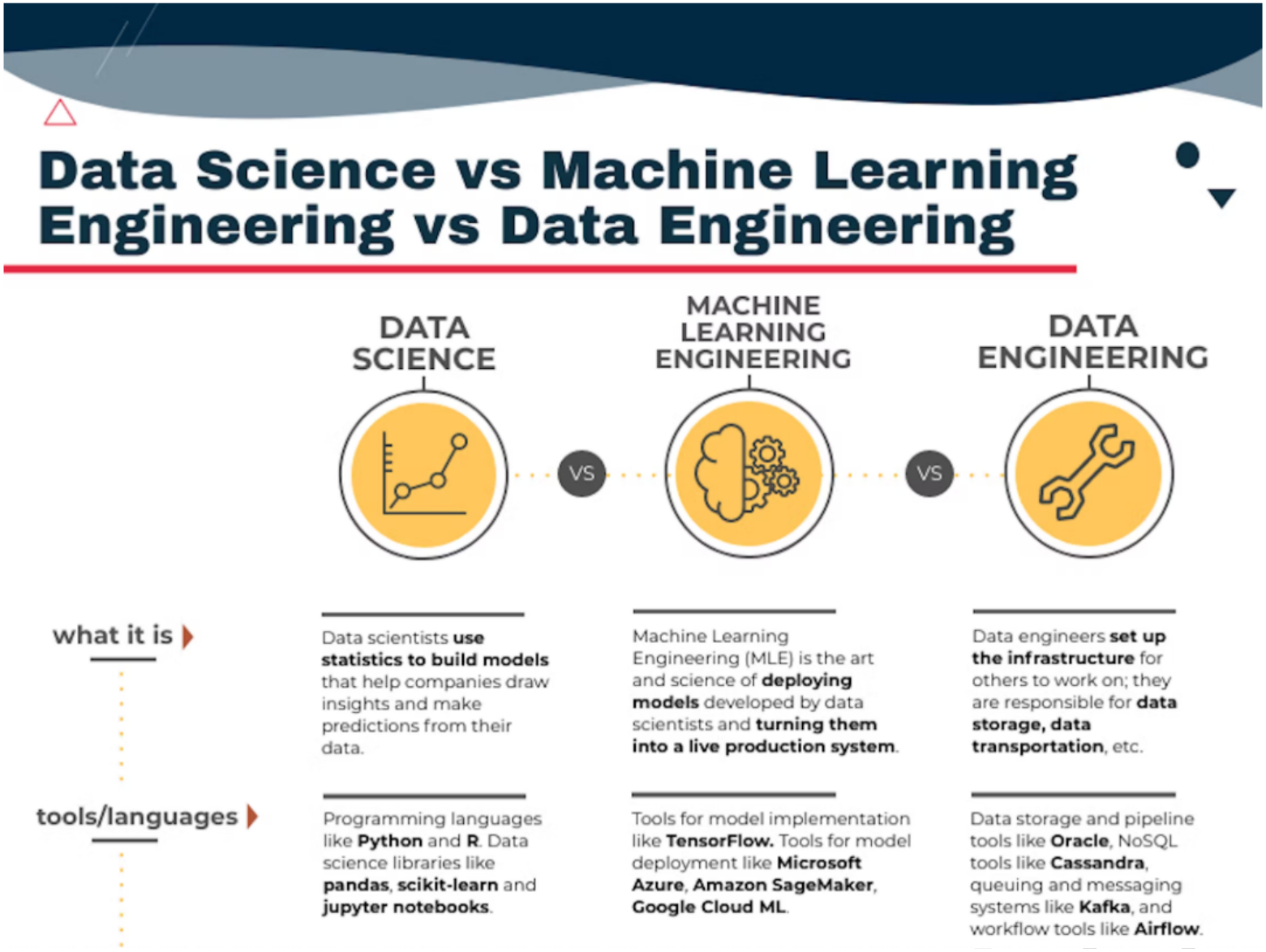
Matemática e Estatística

- Cálculo e Álgebra Linear
  - → Operações matriciais.
  - → Derivada e Integral.
  - → Ponto ótimo em superfície de respostas.

Computação

- Atividades de um cientista de dados
  - → Cloud (GCP, AWS, Azure, etc)
  - → Linguagens de programação: Python, R, SQL.
  - → Versionamento: git.
  - → IDE: VScode, RStudio, PyCharm, spyder...
  - → Etapas de modelagem.

Cientista de Dados



Exemplo

Scenario: A bank wants to detect credit card fraud.

The **data scientist** develops a model that theoretically can detect credit card transaction fraud at a bank.

The **machine learning engineer** would then be responsible for deploying the model in real life and making sure it can handle billions of transactions daily.

The **data engineer** would be responsible for ensuring that all the transaction data that the bank handles is stored properly. If the system needs to process a million transactions a second, the data engineer would build the data pipelines that can convey all that info at the right time to the right parts of the system without any delays or bottlenecks.

Growth Hacking

Alavancar o crescimento com estratégias baseadas em dados

Growth Hacking

Alavancar o crescimento com estratégias baseadas em dados

- Aquisição

Growth Hacking

Alavancar o crescimento com estratégias baseadas em dados

- Aquisição
- Engajamento

Growth Hacking

Alavancar o crescimento com estratégias baseadas em dados

- Aquisição
- Engajamento
- Retenção

Growth Hacking

Alavancar o crescimento com estratégias baseadas em dados

- Aquisição
- Engajamento
- Retenção
- Crescimento

Growth Hacking

Alavancar o crescimento com estratégias baseadas em dados

- Aquisição
- Engajamento
- Retenção
- Crescimento

→ Crescimento da base ativa.

Growth Hacking

Moment A-ha!

Growth Hacking

Moment A-ha!

- Facebook
  - → 7 amigos em 10 dias.

Growth Hacking

Moment A-ha!

- Facebook
  - → 7 amigos em 10 dias.
- Twitter
  - → Seguir 30 pessoas.

Growth Hacking

Moment A-ha!

- Facebook
  - → 7 amigos em 10 dias.
- Twitter
  - → Seguir 30 pessoas.
- Slack
  - → Trocar 2000 mensagens

Growth Hacking

Moment A-ha!

- Facebook
  - → 7 amigos em 10 dias.
- Twitter
  - → Seguir 30 pessoas.
- Slack
  - → Trocar 2000 mensagens
- Netflix
  - → Assistir algo em 90 segundos

Melhoria contínua

Ciência de dados viabiliza adaptação e evolução.

Melhoria contínua

Ciência de dados viabiliza adaptação e evolução.

- Netflix
  - → Assistir algo em 90 segundos

Como melhorar o processo?

Melhoria contínua

Ciência de dados viabiliza adaptação e evolução.

- Netflix
  - → Assistir algo em 90 segundos

Como melhorar o processo?

- → Experimentação.
- → Conclusão a partir de dados.

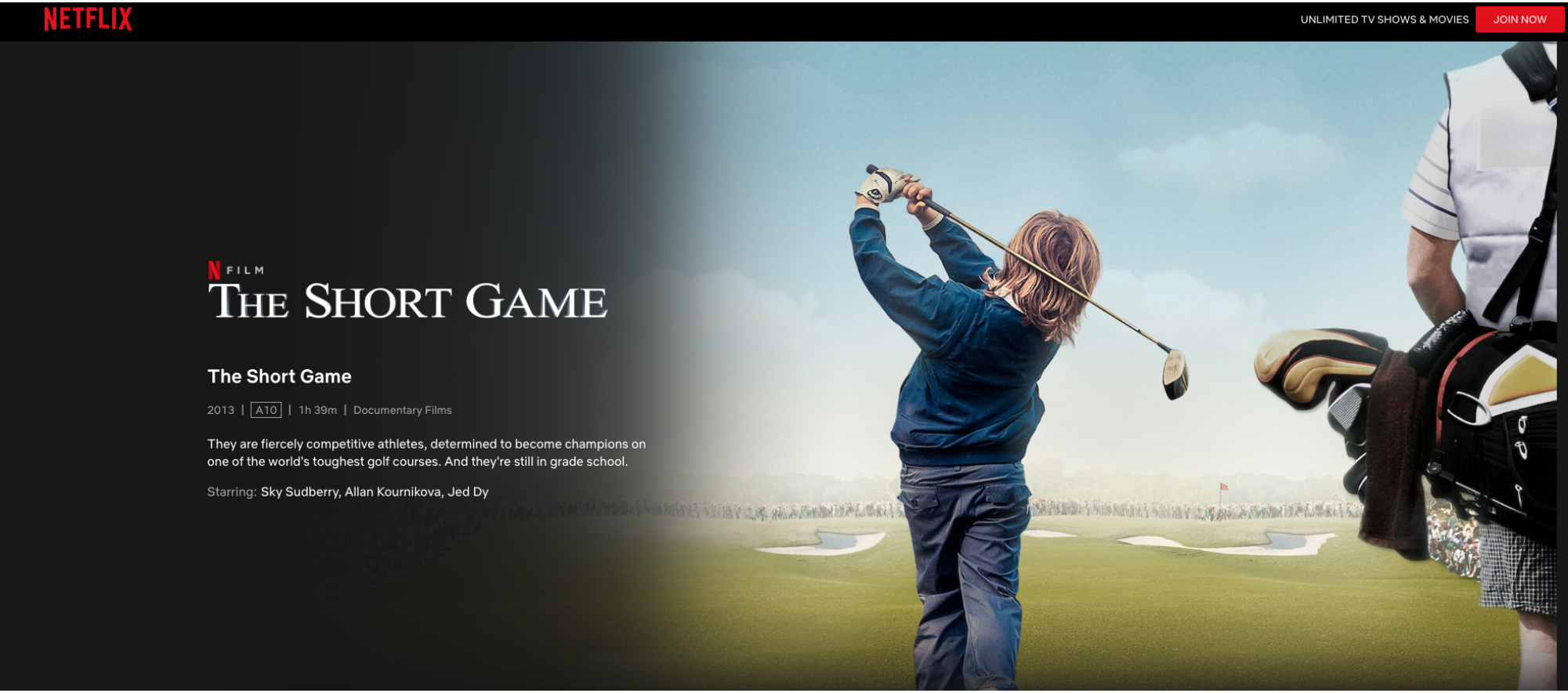
Thumbnail


They are some of the
greatest golfers in the world.
And they are 7 years old.
THE
SHORT
GAME
A FEATURE DOCUMENTARY
www.theshortgamemovie.com

Thumbnail

Cells	Cell 1 (Control)	Cell 2	Cell 3
Box Art			
	Default artwork	14% better take rate	6% better take rate

Thumbnail


NETFLIX
UNLIMITED TV SHOWS & MOVIES
JOIN NOW
FILM
THE SHORT GAME
The Short Game
2013 | A10 | 1h 39m | Documentary Films
They are fiercely competitive athletes, determined to become champions on one of the world's toughest golf courses. And they're still in grade school.
Starring: Sky Sudberry, Allan Kournikova, Jed Dy

Ciência de dados

Transformar dados em informação

- Analisar dados para obter insights
- Identificar padrões, tendencias, correlacoes.
- Entender e contextualizar os dados.

Variável aleatória

Etimologicamente, **Variável aleatória** é uma:

Variável aleatória

Etimologicamente, **Variável aleatória** é uma:

- **Variável**: Possui um valor desconhecido

Variável aleatória

Etimologicamente, **Variável aleatória** é uma:

- **Variável**: Possui um valor desconhecido
- **Aleatória**: Pode possuir valores diferentes, com diferentes probabilidades

Em casos aplicados, nossas "variáveis aleatórias" são as informações observadas.

Tipos de dados

- Quantitativos
  - → Discreta
  - → Contínua
- Qualitativos
  - → Nominal
  - → Ordinal

Quantitativos

Dados quantitativos assumem valores numéricos em sua forma natural (dados numéricos).

- Discretos
  - Assume valores inteiros. Números naturais
  - Ex: Número de "caras" após 3 lançamentos de uma moeda.
- Contínuos
  - Pode assumir qualquer valor no espaço amostral.
  - Ex: Quantidade de água em uma garrafa.

Qualitativos

Dados qualitativos assumem valores categóricos em sua forma natural (categorias, classes)

Não possui significado matemático

- Nominal
  - Não possui hierarquia.
  - Gênero, Nacionalidade.
- Ordinal
  - Possui relação entre valores assumidos.
  - Grau de instrução.
  - Em geral, podemos ter variáveis latentes classificadas como dados ordinais (Ex: avaliação de satisfação)

Dados

Estruturados x Não Estruturados

Estruturados

- Informação organizada
- armazenamento mais leve
  - Tabelas (csv, txt, excel)
  - Bancos de dados relacionais

Não Estruturados

- Estrutura diversificada
- armazenamento mais pesado
  - Imagens
  - Videos
  - Texto

Dados Estruturados

Como os dados estao gravados?
Arquivos e extensões

- .csv (comma-separated-values)
  - → ”,” ”;”
- .txt (text)
  - → tab, espaço
- excel
  - → .xls, .xlsx
  - → tambem suportam csv
- json
  - → chave valor

Dados Tabulares

Forma mais comum de dados estruturados

Tabela: linhas e colunas

Em geral:

- Linhas: observações (ex: cada pessoa)
- Colunas: variaveis (ex: atributos de cada pessoa)

Exemplo

Dados tabulares

- Considere a seguinte tabela:

Ramo	Vendas (R$)	Aluguel (R$)	Nº aluguéis	Tipo
Doces	2.000	300	24	Quiosque
Papelaria	3.000	700	6	Loja
Magica	1.500	600	2	Quiosque
Amendoim	5.000	1000	36	Loja
Pipoca	4.000	800	30	Loja

Dados tabulares

- Considere a seguinte tabela:

Ramo	Vendas (R$)	Aluguel (R$)	Nº aluguéis	Tipo
Doces	2.000	300	24	Quiosque
Papelaria	3.000	700	6	Loja
Magica	1.500	600	2	Quiosque
Amendoim	5.000	1000	36	Loja
Pipoca	4.000	800	30	Loja

Classifique as variáveis da tabela.

Dados tabulares

- Considere a seguinte tabela:

Ramo	Vendas (R$)	Aluguel (R$)	Nº aluguéis	Tipo
Doces	2.000	300	24	Quiosque
Papelaria	3.000	700	6	Loja
Magica	1.500	600	2	Quiosque
Amendoim	5.000	1000	36	Loja
Pipoca	4.000	800	30	Loja

Classifique as variáveis da tabela.
O que são parecem ser as observações?

Dados tabulares

- Considere a seguinte tabela:

Preço (R$)	Tipo Quarto	Quarto Dividido	Capacidade
194	Quarto Privativo	Não	2 pessoas
264	Quarto Privativo	Não	2 pessoas
344	Casa toda	Não	4 pessoas
433	Quarto Privativo	Não	2 pessoas
485	Casa Toda	Não	6 pessoas

Dados tabulares

- Considere a seguinte tabela:

Preço (R$)	Tipo Quarto	Quarto Dividido	Capacidade
194	Quarto Privativo	Não	2 pessoas
264	Quarto Privativo	Não	2 pessoas
344	Casa toda	Não	4 pessoas
433	Quarto Privativo	Não	2 pessoas
485	Casa Toda	Não	6 pessoas

Classifique as variáveis da tabela.
O que são as observações?

Dados Tabulares

Importar e manipular

- R
- python
- excel

Devemos acessar os dados para manipular e trabalhar

Medidas resumo

Não vamos trabalhar com um único valor

- 22

Medidas resumo

Não vamos trabalhar com um único valor

- 22

E sim com vários valores

- $22, 22, 23, 24, 25, 25, 26, 26, 27, 39, 59, 79$

Medidas resumo

Não vamos trabalhar com um único valor

- 22

E sim com vários valores

- $22, 22, 23, 24, 25, 25, 26, 26, 27, 39, 59, 79$

Suponha que a sequencia acima represente uma amostra aleatória contendo a idade de 12 pessoas. Que informações podemos extrair?

Medidas resumo

- Média

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Mediana
  - Valor central

- Moda
  - Valor com maior frequência

- Amplitude
  - Diferença entre o maior valor e o menor valor.

Exercícios

Considerando a idade das 12 pessoas na amostra coletada, calcule: média, moda, mediana, amplitude.

- $22, 22, 23, 24, 25, 25, 26, 26, 27, 39, 59, 79$

Exercícios

Com uma amostra de tamanho 12, é simples calcular na mão. Para lidar com uma grande quantidade de dados, precisamos utilizar ferramentas disponíveis: R, python, VSCode, RStudio, SQL...